View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Asia University Academic Repositories

Review and Analysis of Asia University's 2014 Freshman English Placement Test: *Transition from Version 2.5 to Version 2.6*

Jeff Hull, Jay Brennen and Lindsay Wells, Asia University

Abstract

Asia University's primary assessment instrument for placing first-year students in year-long English classes is the Freshman English Placement Test (FEPT). This article reviews and analyzes the changes made to Version 2.5 of the test that was used in 2013 in order to develop Version 2.6 for the April 2014 administration of the test. The Assessments Committee at the university's Center for English Language Education (CELE) carried out the revision as part of its ongoing efforts to improve the performance and placement accuracy of the test. In order to compare the two versions of the test, we completed standard measurements of test analysis, including measurements of the distribution of scores, means, standard error of measurement, reliability, item discrimination, and test difficulty. Our analysis reveals that although the test continues to perform adequately for its purpose of placing students in English classes and although some of the changes we made resulted in improvements, the revisions in the test resulted in a number of important test measurements having declined. The article concludes that we need to continue to monitor the performance of the FEPT, examine how to improve it, and consider alternatives to the test.

Introduction

The starting point of the Assessments Committee's work on creating Version 2.6 of the FEPT was its review of the performance of Version 2.5 in 2013. Hull and Brennan concluded that the new word discrimination part the committee had developed for Version 2.5 did not perform as well as the committee had hoped and had, in fact, performed worse than the word discrimination part it was developed to replace (2014, p. 57). Since that particular part of the test had historically been one of its poorer performing parts (Hull, 2012, p. 10) and because the committee had reservations about using word discrimination as a method of discriminating among student abilities (Hull and Brennan, 2014, p. 58), the committee decided to remove it altogether and include a new monolog part. The committee also identified a number of test items in the remaining three listening tasks of the test that could be edited or replaced because they performed poorly in identifying students at one or another proficiency level for placement purposes. The committee then decided to create a pilot test to screen items for the new monolog part and as replacements for weaker performing items in the test.

To help clarify the work the Assessments Committee carried out in 2013, Table 1 identifies the modifications made in the FEPT in recent years. The reader can find a more detailed account of the history of the test in Hull and Brennan's article (2014, pp. 35-37).

Table 1: Summary of FEPT Development in Recent Years

Test/Year	Number of Items	Time	Sections/Parts
Version 2.3 2007	98	54:00	Listening Section: Parts 1-5 Vocabulary, Grammar and Reading Section: Parts 6-8
Version 2.4 2012	75	39:30	Listening Section: Parts 1-4 Vocabulary, Grammar and Reading Section: Parts 5-7
Pilot Test 2012	11 55	---	Listening Section: Alternative 1 Part 1 Word Discrimination Alternative 2 Part 1 Word Discrimination
Version 2.5 2013	75	39:30	Listening Section: Parts 1-4 (with 11 items from Alternative 2 of Part 1 to make the new Part 1 of the test) Vocabulary, Grammar and Reading Section: Parts 5-7
Pilot Test 2013	24	---	Listening Section: Alternative Part 2 Picture Identification (7 items) Alternative Question and Answer items (2 items) Alternative Dialogs (3 items) New Monologs Part (12 items)
Version 2.6 2014	72	40:00	Listening Section: Word Discrimination, Alternative 2 Part 1, is eliminated. Picture Identification, Part 2, is moved to Part 1 (includes 2 new items) Question and Answer, Part 3, is moved to Part 2 (includes 3 edited items) Dialogs, Part 4, is moved to Part 3 (includes 2 edited items) Monologs becomes the new Part 4 (includes 6 new items) Vocabulary, Grammar and Reading Section: Parts 5-7

This paper will review and analyze how Version 2.6 of the test performed in its first year of use to determine whether the new monolog part and other new replacement items resulted in improvement. Additionally, the paper will compare the number of complete scores obtained for students at the end of the 2013 to 2014 academic year with previous years to follow up on an issue the Assessments Committee has examined in recent years regarding how effectively the test provides the Academic Office with the scores it needs at the end of students' first academic year to place them in English classes in their second year (Hull and Brennan, 2014, pp. 56-57). Finally, the paper will consider directions for improving or replacing the current FEPT for the future.

I. From Version 2.5 to Version 2.6

To develop new material for Version 2.6 of the test, the Assessments Committee created a 24-item pilot test which included a new monolog part and potential replacements for some of the poor performing items in Version 2.5. The committee then recorded the Listening section and administered the test to first year International Relations students since they take the TOEIC instead of the FEPT for placement in English classes.

After the results of the pilot test were analyzed, six of 12 items were selected to create the new monolog part of Version 2.6 (Part 4), two new items were added to the picture identification part (Part 1), two items were edited in the question and answer part (Part 2), and three items were edited in the dialog part (Part 3). The committee also re-recorded the entire Listening section of the test since the audio for Version 2.5 was of inconsistent quality. Other than these changes, the test was kept the same as Version 2.5 so that when the committee analyzed the results it could focus on how the limited number of revisions had affected the overall performance of the test.

To maintain a record of how the Assessments Committee carried out its revision of the FEPT and to serve as a guide to future CELE Assessment Committees, we will provide a brief account of the development of the pilot test together with an analysis of the results of the test to create Version 2.6.

A. New Items

One of the major challenges in creating the 2013 Pilot was compensating for the elimination of Part 1: Word Discrimination in Version 2.5. This part of the test contained 11 items. The prompt for each item consisted of a single word, repeated once. Test takers were tasked with choosing the correct word from a set of four options. Due to the minimal spoken text in this part, the audio recording for this part, including instructions, was only two minutes and 43 seconds (2:43) long. In comparison, the overall length of the test audio was 20 minutes.

Because the FEPT would need to be administered within a 45-minute class period, it was not feasible to significantly lengthen the audio recording. This posed logistical issues as all other parts within the listening portion of the test featured

significantly longer spoken texts, meaning that fewer of these items could be included in the available time. In general, as test length decreases, so does reliability (Hughes, 2009, p. 36); consequently, it was important to reduce the total number of items in the test as little as possible. To solve this problem, the committee chose to concentrate on developing two task types, one of which was relatively short, and one that featured longer texts that were each associated with multiple items.

The first task type was picture identification. These items, which have been included on the FEPT since its inception, consist of a picture and four short statements. The test taker must select the statement that best describes the picture.
Such items are both simple and short, running less than 30 seconds in length each. The second task type was monologs. These had been included in past versions of the FEPT but were eventually discontinued due to their difficulty. The committee felt that perhaps that difficulty was not due to the task itself, but to the content of the individual monologs. In addition, part of listening in a foreign language consists of comprehending extended speech, and so a monolog may add to the construct validity of the test. Each monolog was relatively long, with a time of one minute (1:00) to one minute 40 seconds (1:40). Instructions would need to be included as well, adding even more time to the proposed new section. Following the model of the TOEIC, however, each monologue could be tied to three items, bringing down the time per item significantly. Both item types, then, were deemed adequately efficient.

The committee then needed to determine the content of the new items. In Version 2.5, the committee had noted that items tended to test easier language points, and so more difficult items were needed. In writing the items, then, the item writer primarily referred to the textbooks designated for the highest levels of Freshman English: Four Corners 2 and Four Corners 3. Some easier items based on in-house materials and Four Corners 1 were also included, mainly in order to add breadth to the topics in the existing test. Whenever possible, the item writer also took care to use scenarios that were authentic, and that students might encounter in real life.

A total of five new items were created for the picture identification part. Pilot item 1 tested phrases related to classroom English. Item 2 required knowledge of the present perfect. Item 3 tested vocabulary related to business, school, and leisure activities.

Item 4 centered on household chores. Finally, item 5 focused on the use of "too" and "enough" as intensifiers. Images for all items except number 3 were created by the Assessment Committee; the remaining image was a creative commons photograph sourced from the internet.

In addition to the five picture identification items, 12 monolog items were created for the pilot test. These would come at the end of the listening section, so they should be the most difficult. The monolog task consisted of longer texts said by a single speaker, and was designed to test students' comprehension of extended speech. Previous monologs featured advanced and irrelevant content for the target population. For example, in one item from Version 2.3, the speaker gave a short talk about James Joyce. Such a literary topic could prove frustrating for the incoming students; moreover, it does not relate to the content taught in Freshman English, so students' ability to answer this type of question would mean little in the context of Asia University's English curriculum. To avoid a similar situation in Version 2.6 of the FEPT, the item writer chose to create scenarios that students might encounter in the course, while integrating linguistic structures and vocabulary words from the course textbooks themselves. Monolog 1, for example, consisted of a phone message in which a man talks about a job opening. In Monolog 2, a teacher speaks to a new Freshmen English class. Monolog 3 is a woman recounting an incident on the train. Finally, Monolog 4 features a person giving a speech about studying abroad. Each monolog had three corresponding items, which featured language points covered in the upper-level Freshmen English textbooks.

In all, the committee piloted 17 new items. They did so knowing that 17 items would be too long for the allotted space. However, the committee preferred to have a larger pool of items to choose from when the time came to decide which ones to include in Version 2.6. In addition to these new items, the pilot also included a number of revised items. These will be briefly described in the following section.

B. Revised Items

Six items from the previous two versions of the FEPT, Versions 2.4 and 2.5, were identified as in need of revision based on their item discrimination index values. The item discrimination index measures how well an item separates the top performing test takers

from the lowest performing test takers. In a previous analysis of the FEPT, Hull identified a discrimination index value of below .20 as the level at which items should be more closely examined for replacement or removal from the test (2012a, p. 6). Over the past four years, the items in question consistently had discrimination indices below .20. These values can be seen in Table 2. The top row indicates the item numbers for the 2013 Pilot test; the second row indicates those for Version 2.4 of the FEPT, and the fifth row indicates the items for Version 2.5.We will now briefly explain the proposed changes for each of these items. While making these changes, the writer of these items drew on widely-held item writing principles (Haladyna and Downing, 1989; Haladyna, Downing, and Rodriguez, 2002). For the sake of clarity, we will use the item numbering from the 2013 Pilot to refer to individual items.

Table 2: Discrimination Index Values for Pilot Items 6-12, Before Revision

2013-Pilot	ITEM #	Item 6	Item 7	Item 8	Item 9	Item 10	Item 12
Year	ITEM #	Item 12	Item 13	Item 17	Item 21	Item 32	Item 38
2013-1	DISC INDEX	0.10	0.17	0.03	0.11	0.15	0.15
2012-1	DISC INDEX	0.12	0.11	0.13	0.10	0.24	0.16
	ITEM #	Item 13	Item 14	Item 21	Item 25	Item 38	Item 44
2011-1	DISC INDEX	0.11	0.25	0.18	0.08	0.10	0.09
2010-1	DISC INDEX	0.24	0.10	0.09	0.19	0.14	0.19

Items 6 and 7 belong to the picture identification portion of the test. Item 12 was based on a picture of students playing badminton in a gymnasium. The key for this item was option C, “They’re playing in the gymnasium.” The featured vocabulary here is high frequency and typical of school settings, making the difficulty of this option relatively low. On the other hand, two of the three distractors relied heavily on words relating specifically to badminton (“net”, “racket”, and “bird”). “Net” and “racket” do have cognates in Japanese, but “bird” (as in “shuttlecock”) does not. Unless a test taker had knowledge of badminton terminology in English, which is unlikely, references to birds would be nonsensical, not distracting. The third option, “They’ve lost their mittens”, is

also problematic because it is the only answer not in the present continuous. This muddles the linguistic focus of the item. The proposed change, then, was to make the grammar consistent across all four distractors, and to have the difficulty hinge on the meaning of common verbs ("standing", "practicing", "waiting", and "playing"), thus making the item more uniform in both structure and relative difficulty.

Item 7 posed a similar problem. The key for this item, which centered on a train waiting in a station, was "The doors are open." Both the vocabulary and the grammar in this statement are basic, and likely to be understood by almost all students. The vocabulary in the other options were either very basic and easily eliminated ("plane") or a more sophisticated play on words ("stationary", which sounds like "station"). As in item 7, the item writer felt that the options should be more grammatically and contextually uniform, so simpler statements referring to trains and train stations were used. The item writer felt that, as a result, the key would be a less obvious choice.

Items 8 and 9 belong to the question and answer part of the exam. In this task type, students hear a question and must select the best response. The question for item 8 was originally "When is your next class?" The original key was "At 10:40". The other two distractors were "For Freshman English", and "In Building 2". These three options are strong in that they are uniform in sentence structure, of a similar difficulty level, and tempting. Still, the item's poor performance indicated that there was room for improvement. After giving the item a second look, the item writer noted that "In Building 2" may not function well as a distractor because the word "building" has a cognate in Japanese. This would be easy for many students to rule out. In addition, two of the three options featured cardinal numbers ("two", "ten" and "forty"); this could make the third option stand out as strange. If "In Building 2" used an ordinal number such as "second" or "fourth" instead, the options would be more varied. An ordinal number could also be tempting to students who associated the ordinal with a date, such as "December fourth". For these reasons the item writer changed "In Building 2" to "On the third floor." The question was also changed to "When is the test?" The other two options remained the same.

For Item 9, the prompt was "How can I get in touch with you?" In the original item, the key and one of the distractors hung on the idea of contacting someone. The

second distractor, however, was “Don’t touch me.” This stood out as an odd (and potentially distressing) thing to say. The item writer decided to remove this distractor from the test. The remaining distractor, “Yes, you can anytime,” was tempting because it was a yes/no question, and also because it referenced time. The item writer then shortened this option to “Yes, you can”, and added an additional option referencing time. In effect, then, the stronger distractor was split into two. After these changes, all options for this item were now reasonable and tempting.

Items 10 and 12 are dialog items. In this part of the test, students listen to a dialog and then answer either one or two comprehension questions. The dialog for Item 10 took place between a teacher and a student who was absent for class, missing a quiz. They talk about why the student was absent, and the dialog ends with the teacher saying, “Well, you’ll have to take the quiz later. See me after class.” Test takers must then identify what the student in the dialog should do, with the correct answer being, “See the teacher later.” This item proved to be problematic on a number of levels. First, the quality of the audio was poor. Second, half of the options pertained to the student taking the quiz, and the other two pertained to speaking with the teacher, meaning that students may be able to quickly eliminate half of the answers, effectively turning the item into a two-option item. Third, and most importantly, one of the distractors was, “See me after class.” The identity of “me” in this distractor is inherently unclear, making the option misleading, confusing, and potentially a second key. The item writer eliminated this option as well as one of the options referencing the quiz. The writer then created two new options referencing potential excuses for the student being absent (sickness and missing the train.) As a result, the distractors were more well-rounded, tempting, and clearly incorrect.

The dialog for Item 12 took place between two students. One student complains about having to take the train to school, and the other student suggests moving into a dormitory. The test taker must then identify what the second student thinks the first student should do. Because dormitories are typically located on campus, the key is “Move closer to school.” To get the answer correct, a student would have to know that a dormitory is usually on campus. Many Japanese university students live at home, however, and on-campus dormitories are less common in Japan than in other countries. It is thus possible that a test taker who recognized the word but did not realize that it

referred to an on-campus location could get the item wrong. On the other hand, students who do live in a dormitory would be very familiar with the location of that kind of building. For this reason, the item's difficulty could lie to a large extent on topical knowledge rather than language knowledge, making it inappropriate for a language test (Bachman and Palmer, 1996). To fix this, the item writer changed the phrase "moving into the dormitory" to "finding a place to live near here." The original meaning is still implied, but the key phrase no longer depends on topical knowledge. As this revision involved changing the dialog itself, Item 39 from Version 2.5, which was based on the same dialog, was included in the pilot as well as item 11. For this reason, while only six items from Version 2.5 were revised, a total of seven were included in the pilot test.

C. Compilation and Administration of the Pilot Test

Once the proposed items were written, the Assessment Committee compiled the items to be piloted into a single test booklet. The booklet consisted of four parts: Picture Identification, Question and Answer, Dialogs, and Monologs. A short survey was included at the end of the pilot. This survey was used to determine whether students had any problems with understanding instructions or particular items. The committee felt the students' comments were generally unhelpful and consequently the student survey will not be included in future pilots. The Assessments Committee discovered Student Assistants feedback uncovered issues with the pilot and final version of the FEPTs most effectively. Student Assistants work in CELE on a daily basis and typically possess a high proficiency in English. Therefore, they are readily available to assist Assessment Committee members with reviewing and analyzing the FEPT from a student test taker perspective. Finally, a complete new audio file was recorded for the entire FEPT, including revised items. The new audio was a great improvement from the patchwork of audio that had been spliced together over the years. Essentially each time the FEPT had been revised over the years and a new audio item was added this new audio item would be recorded and inserted into the original recording. Due to so many changes over the years there was quite a bit of inconsistency in the audio quality in general. Consequently the audio quality, volume and general clarity of the audio varied from one listening item to the next in some cases. Moreover, the committee had concern this was providing

misleading results in our review and analysis of Listening section items in general. The new audio is very clear now and we are quite happy with the consistency and clarity of all listening items.

D. Results of the Pilot Test

The pilot results were analyzed in terms of discrimination index and facility values. These are reported in Table 3. Discrimination Index values of .20 or above, the level the committee has considered acceptable for the FEPT, are indicated in italics.

Table 3: Item Discrimination and Facility Values for Piloted Items

PILOT ITEM	1	2	3	4	5	6	7	8	9	10	11	12
Disc Index	0.08	***0.20***	-0.02	***0.35***	***0.27***	0.08	-0.06	***0.37***	***0.43***	***0.24***	***0.24***	***0.39***
Fac Values	0.28	0.36	0.94	0.50	0.26	0.89	0.95	0.77	0.59	0.51	0.70	0.29

PILOT ITEM	13	14	15	16	17	18	19	20	21	22	23	24
Disc Index	***0.20***	***0.43***	***0.20***	***0.24***	***0.43***	***0.25***	0.12	***0.25***	***0.31***	0.18	0.06	***0.25***
Fac Values	0.62	0.31	0.28	0.70	0.48	0.45	0.53	0.50	0.42	0.47	0.23	0.56

The 17 items that reached the .20 level or above included two of the six revised items, as well as three of the new picture identification items and nine of the twelve monolog items. Of the successful monolog items, three corresponded with Monolog 1, three with Monolog 2, two with Monolog 3, and one with Monolog 4. Table 4 shows the breakdown of the monolog items.

Table 4: Monolog Items with Favorable Discrimination Index Values

Monolog	Average Discrimination Index Value	Items
Monolog 1	.27	***13***, ***14***, ***15***
Monolog 2	.31	***16, 17, 18***
Monolog 3	.23	19, ***20***, ***21***
Monolog 4	.16	22, 23, ***24***

Due to their favorable statistics, the committee decided to adopt both of the successful picture identification items as well as all of the successful revised items. The committee also selected the two stronger performing monologs, Monolog 1 and Monolog 2.

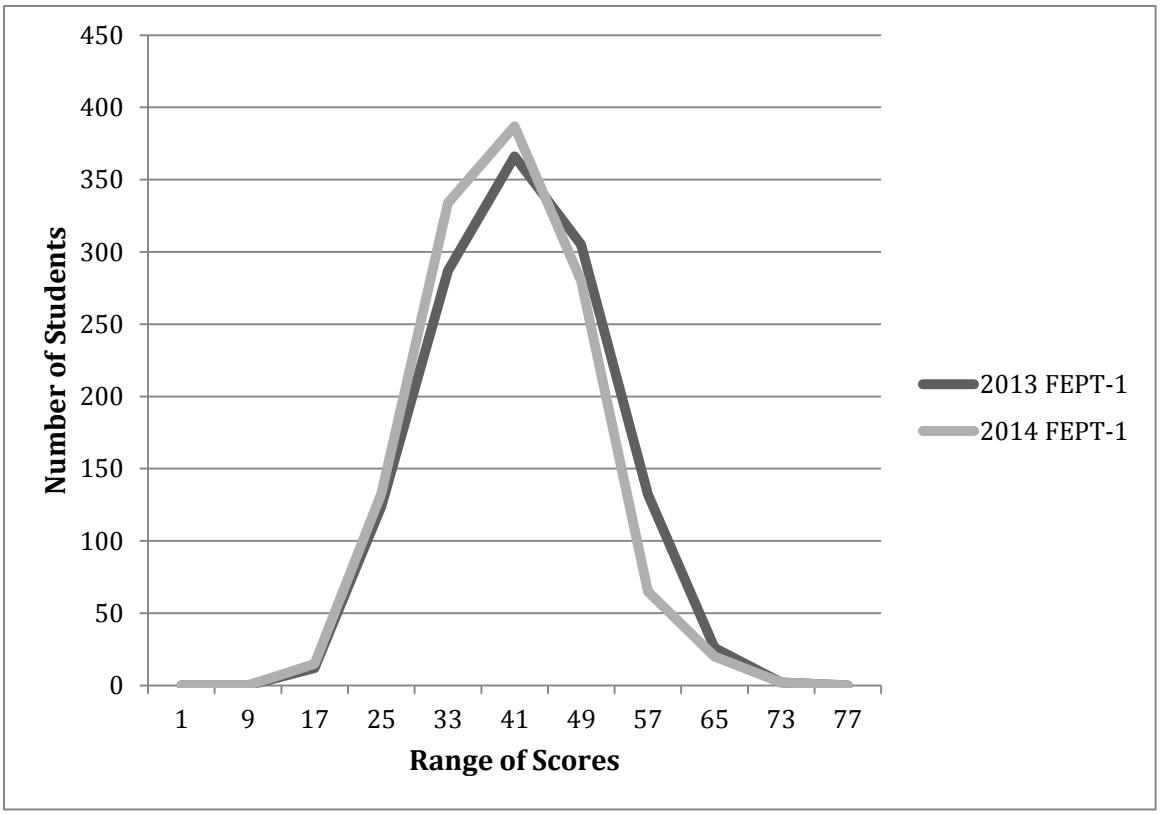
Overall, the monologs had relatively high facility values, meaning that a large proportion of test takers were able to answer them correctly. There was some concern that the items might be too easy, but the Assessments Committee felt that this was most likely due to the fact that a typically more motivated and well-performing population took the pilot test. In addition, the pilot group took the test near the end of the academic year, after having benefitted from almost a whole year of English language instruction. The committee assumed that in comparison, incoming students from other departments would find the test more difficult, resulting in lower facility values. For these reasons, the committee felt that items 4, 5, 8, 9, 10, 12, 13, 14, 15, 16, 17, and 18, could be positive additions to the FEPT, and that these items would therefore be written into Version 2.6. We will now look at the results of Version 2.6.

II. Analysis of Version 2.6 FEPT

A. Distribution of Scores

As can be seen in Figure 1, the distribution of scores for the April 2014 FEPT is very close to the distribution for April 2013. Most of the 2014 scores are inside the middle 60 percent of the distribution, similar to the 2013 results, and the graph is not noticeably leaning to the left or right but instead has a symmetrical shape. Although the issue of test difficulty will be addressed later in the paper, this is one indication that the level of difficulty of the test was appropriate to our test population.

Figure 1: Distribution of Scores, April 2013 and 2014 FEPT



One of the main purposes of a placement test is to divide students into different class levels (Harris, 1969, pp. 125-126). The measurement of standard deviation is one of the primary ways of determining how well a placement test achieves this goal. The higher the standard deviation, the more widely scores are distributed and the easier it is to divide students into different classes. Although Table 5 shows that the standard deviation for the 2014 FEPT decreased slightly compared to 2013, the lower number is not statistically significant and does not indicate a decline in the test's ability to place students into an appropriate level Freshman English class. Taking into consideration the fact that Version 2.6 has three fewer items than Version 2.5, it represents about the same range of distribution relative to the total number of items. On the other hand, it does not represent the improvement the committee had hoped to make since the performance of the 2013 version of the test had declined in this area compared to the 2012 version.

Although the drop in the mean by nearly two points may at first appear significant, it actually represents only a small increase in the level of difficulty of the test

after taking into consideration the fact that Version 2.6 has three fewer items. The fact that there was some movement in the direction of making the overall test more difficult is actually one of the positive changes the committee was hoping for and a subject which we will address in Section D of the paper dealing with test difficulty.

Table 5: Details FEPT Test Measurements, 2012-2014

FEPT Test	Number of Items	Number of Examinees	Mean	Std. Error of Measurement	Std. Deviation
April 2012	75	1178	39.2	3.9	10.5
April 2013	75	1254	38.1	3.9	9.7
April 2014	72	1236	36.3	3.8	9.1

B. Reliability

One of the key measures of how well a test performs is its reliability, the ability of the test to provide consistent results with a particular test population. In order to have confidence in an entrance test, we need to confirm that the test provides approximately the same results from one year of first year students to next. We calculated two very standard measurements of reliability, Cronbach's Alpha and Kuder Richardson 21, so that we would be able to compare the 2014 test results with the results from previous years.

The index that is obtained from a reliability measure ranges from zero to 1.00. An index of 1.0 represents perfect reliability, in other words a test that gives exactly the same results for a particular set of examinees each time the test is administered. On the other hand, an index of zero represents the complete absence of reliability. In such a case, the scores that examinees get on one administration of the test vary so greatly from the scores they get on the next administration of the test that the two sets of results seem completely unconnected to each other.

Agreement among language testing experts about what constitutes an acceptable level of reliability for placement tests has not been reached. However, one well-recognized expert, Arthur Hughes, suggests that a figure between .80 and .89 is desirable for listening comprehension tests and between .90 and .99 for vocabulary, structure, and reading tests (2009, p. 39). It would be unrealistic, however, to expect that a test like the

FEPT, which has been produced by a small number of CELE teachers with an interest in testing but limited expertise, time and resources, to reach that high a level of reliability. Harris refers to tests like the FEPT which have not been produced by an independent professional testing organization as "homemade tests," and as such they would more typically have measurements of reliability in the .70s or .80s (1969, p. 17).

Taking into consideration both Hughes's and Harris's comments on reliability, the level of reliability that the FEPT has achieved and maintained is acceptable. Table 6 shows that the measures for both Cronbach's Alpha and KR 21 have historically reached the low to mid .80's, and this is true for the 2014 administration of the test as well. Version 2.6 of the FEPT still satisfies its function as our primary placement instrument for separating students into four or five broad levels of ability so that they can be placed in Freshman English classes. On the other hand, Table 6 also shows a slight decline in reliability for 2014 compared to 2013 and the years before that. Here again, because of the decline in reliability from 2012 to 2013, the drop in 2014, although not statistically significant, represents movement in the wrong direction that the Assessments Committee would like to remedy.

Table 6: Measurements of Reliability for the FEPT, 2008-2014

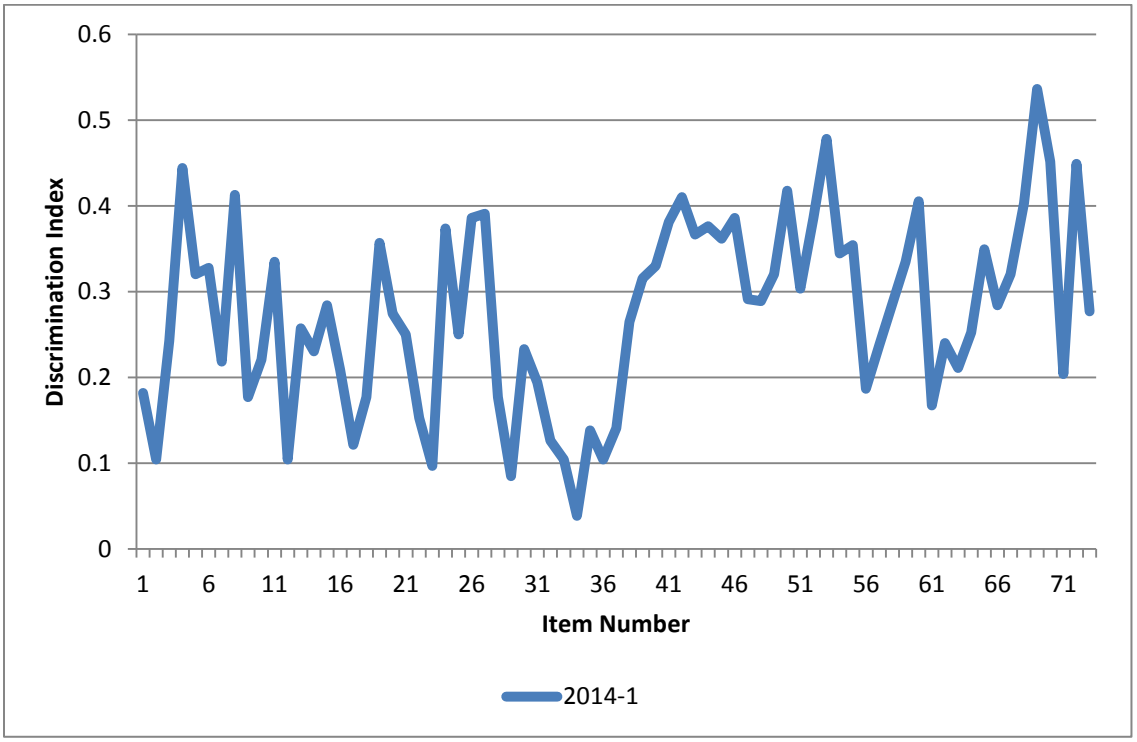
FEPT Test	Version of FEPT	Number of Items	Cronbach's alpa	KR21
April 2010	2.3	98	.86	.84
April 2011	2.3	98	.85	.83
April 2012	2.4	75	.86	.84
April 2013	2.5	75	.84	.81
April 2014	2.6	72	.83	.80

C. Item Discrimination

Another critical measurement of test performance concerns how well each item in the test separates students of higher or lower ability. Figure 2 shows a visual representation of this measurement of item discrimination values for Version 2.6. A number of items in the first half of the test, the Listening section, are below a discrimination index value of .20, which is the minimum standard the Assessments Committee has used to more closely examine whether an item should be replaced or

removed. This has been noted in past reviews of the test and has been a primary focus of the Assessments Committee's attention in recent years (Hull and Brennan, 2014, p.51-52).

Figure 2: Item Discrimination for the FEPT 2014



As shown in Table 7, if we average the discrimination index values for the Listening section of the test and compare that figure with the average for the Vocabulary, Grammar and Reading section of the test we can see this difference in performance even more clearly. The average discrimination value for the Listening section clears the .20 standard but falls significantly below the average of the Vocabulary, Grammar and Reading section of the test. Both sections dropped slightly compared to the previous year, but it is difficult to identify what to attribute this to. Both sections declined by the same amount despite the fact that only the Listening section was changed from the previous years. It could simply be a result of slight but statistically insignificant variations from one year to the next.

Table 7: Average Discrimination Index by Section and Overall

Test	Listening	Vocabulary, Grammar, and Reading	Overall Average DI
April 2010	.24	.30	.27
April 2011	.23	.30	.26
April 2012	.27	.36	.31
April 2013	.24	.35	.29
April 2014	.22	.33	.28

Table 8 gives us a clearer picture of how the test performed in the area of item discrimination. For Table 8, we isolated the specific items in the test that we changed to create Version 2.6. The most significant change made to the test was the removal of the 11-item word discrimination part, which, as noted previously, had been one of the weaker performing parts, particularly in 2013 after the Assessments Committee rewrote it, and the subsequent addition of a 6-item monolog part and two new items to the picture identification part of the test. The third column of Table 8 shows that, unfortunately, the new items performed worse than the former 11-item word-discrimination part that was removed and well below the .20 minimum standard. On the other hand, as shown in Column 4, the newly edited items in Version 2.6 appear to have improved in their ability to differentiate among student ability levels compared to the previous year although those items, too, fall slightly below the .20 standard.

Table 8: Average Discrimination Index for the FEPT, 2010-2014

FEPT Test	Number of Items	Item Disc Ave for Word Discrimination (Part 1)	Item Disc Ave for 5 Edited Items	Item Disc Ave for All Items
April 2010	98	0.25	X	0.27
April 2011	98	0.25	X	0.26
April 2012	75	0.25	.18	0.31
April 2013	75	0.18	.14	0.29
April 2014	72	**Eliminated Word Discrimination Part But Added 8 new items** **.15**	.18	0.28

Measuring the discrimination indexes for the individual parts of the test reveals even more detail about the performance of the new items and the edited items added to the test. Table 9 shows that the two new items added to Part 1: Picture Identification and two items edited in Part 2: Question and Answer performed well and helped maintain the acceptable discrimination indexes of those two parts. On the other hand, the table also shows that the three edited items in Part 3: Dialogs resulted in a noticeable decline in the average discrimination index for that part. More importantly, the new Part 4: Monologs obtained the lowest discrimination index values, even lower than the previous monolog part of Version 2.3 administered in 2010 and 2011, which was removed from the test because of its poor performance. The reason why the monolog part of the test performed poorly in 2014 (Part 5) will be addressed in the next part of the paper that deals with test difficulty.

Table 9: Discrimination Index by Part for Section 1

	Section 1: Listening				
TEST	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**
April 2010	.25	.22	.24	.29	.16
April 2011	.25	.21	.22	.28	.14
April 2012	.25	.25	.25	.32	Removed
April 2013	**.18**	.26	.22	.29	Removed
April 2014	**Removed**	**Part 1 2 new items .26**	**Part 2 2 edited items .24**	**Part 3 3 edited items .24**	**Part 4 6 (all) new items .11**

One final note to make here is in regard to Table 10. Table 10 shows the Vocabulary, Grammar and Reading section, which the committee did not make any changes to, had discrimination index values that were entirely consistent with the strong performance of that part of the test in the past.

Table 10: Discrimination Index by Part for Section 2

	Section 2: Vocabulary, Grammar and Reading		
TEST	**Part 6**	**Part 7**	**Part 8**
April 2010	.32	.28	.28
April 2011	.33	.26	.31
April 2012	.38	.29	.43
April 2013	.36	.29	.42
April 2014	**Part 5** **.35**	**Part 6** **.28**	**Part 7** **.34**

D. Test Difficulty

The average score on the test for 2014 as indicated in Table 11 is right at the 50% mark, the level of difficulty which is generally considered ideal (Brown and Hudson, 2002, p. 33). An average score considerably above 50% would mean that the test was too easy with too many of the students answering the questions correctly. Such a test would not discriminate among all students sufficiently for placement purposes. It might effectively separate the students at the bottom level of proficiency from the rest of the test population but not those at the middle from the top third. There would be a similar problem if the average score was significantly below the 50% mark and therefore too difficult. The table shows that the test has moved still closer to the 50% mark from 2013 to 2014, a difference that is not statistically significant but welcome nevertheless since it is movement in the desired direction.

However, the table also shows that the two sections of the test continue to be imbalanced in their level of difficulty, with the Vocabulary, Grammar and Reading section being noticeably easier than the Listening section. Ideally, the two sections should be closer in their level of difficulty. The imbalance was a result of the committee's work in 2011 to modify the test so that it could be administered in less time without compromising its overall performance (Hull, 2012, pp. 1-2). To achieve that, a number of items in the Vocabulary, Grammar and Reading section that had low discrimination index values were removed from that section. The primary reason those items had low discrimination values was because they were too difficult for the test population which

resulted in very similar numbers of students at the top and bottom of the scoring distribution answering correctly. That suggests that the students may have simply been guessing at the answers. The outcome of modifying the test for 2011 was both good and bad for the performance of the test. It improved the ability of the test, particularly the Listening section, to differentiate among student ability levels. But it also resulted in the Vocabulary, Grammar and Reading section of the test becoming noticeably easier.

Both test characteristics, the appropriate level of difficulty and the ability to discriminate effectively among student ability levels, are important. However, when both cannot be fully achieved and the committee has to concentrate more attention on one or the other, the ability to divide the students accurately should receive greater attention than balancing the level of difficulty across different sections of the test. In line with that reasoning, the committee has focused more of its time after 2012 on an attempt to boost the comparatively lower discrimination ability of the Listening section than on balancing the difficulty levels of the two sections of the test. Once the committee makes some improvement in the ability of the Listening section to discriminate among students, it can redirect its attention to balancing the difficulty levels of the two sections. In the meantime, erring on the side of having a test section that is slightly easier than is optimally desired is better than frustrating students with a test section that is beyond their ability.

Table 11: Average Scores by Section and Overall (reported as percent correct)

Test	Listening	Vocabulary, Grammar, and Reading	Overall Average Score
April 2010	48%	51%	49%
April 2011	47%	51%	49%
April 2012	48%	57%	52%
April 2013	45%	57%	51%
April 2014	45%	56%	50%

Measuring the facility values, the standard level of difficulty measure, for the individual parts of the test that were changed reveals more detail about how the new items and the edited items influenced the difficulty level of the test. The facility values for the edited items, as can be seen in Table 12, did not affect the test in any significant

way. On the other hand, removing the word discrimination part and adding new items to the test clearly did. The increase of about .16, or 16%, in difficulty level by itself is not necessarily a problem. But that degree of increase in difficulty together with a very low discrimination index average for those items as seen in Table 8 is certainly an undesirable combination.

Table 12: Average Facility Value for the FEPT, 2010-2014

FEPT Test	Number of Items	Fac Val Ave for Word Discrimination (Part 1)	Fac Ave for Edited Items	Fac Ave for All Items
April 2010	98	0.48	X	.49
April 2011	98	0.47	X	.49
April 2012	75	0.50	.34	.52
April 2013	75	0.45	.34	.51
April 2014	75	***Eliminated word discrimination part but added 8 new items*** ***.29***	.36	.50

An examination of the facility values for the different parts of the test reveals that the major contributor to that increased difficulty level was Part 4, the newly created items for the monolog part. Table 13 shows that not only was there a .16 increase in the difficulty level compared to the word discrimination part that was removed but that it was nearly a 10 percent greater level of difficulty than the previous monolog part which itself had been removed from the test in 2012 for being too difficult for this test population and consequently ineffective at discriminating among student ability levels (Hull, 2012, p.7). On the other hand, the two items edited in Part 2: Question and Answer resulted in that part becoming easier than in previous years but not to the extent that Part 4 had contributed to the increase in the difficulty level overall of the Listening section.

One point to note here about the difficulty levels of the various parts of the test is a concern the Assessments Committee has had to follow the intention of the original designers of the test to have examinees move from easier to more difficult items as they completed each of the two different sections of test (Forster & Kerney, 1997, p. 145). This is a standard sequence recommended for tests of language ability (Bachman, 1990, pp. 120-121). With the removal of the word discrimination part, the committee came

closer to achieving that progression in the Listening section than it had in the past and certainly ended the undesirable progression of having examinees begin the test with one of the most difficult parts rather than the easiest. On the other hand, with the decrease in difficulty for the second part of the test as a result of editing a few of the items, the committee did not fully achieve its goal here.

Table 13: Average Scores by Part for Section 1 (reported as percent correct)

	Section 1: Listening				
TEST	**Part 1**	**Part 2**	**Part 3**	**Part 4**	**Part 5**
April 2010	48.8%	55.7%	50.3%	46.6%	37.4%
April 2011	47.3%	53.9%	50%	45.2%	37%
April 2012	49.5%	51.5%	50%	45.1%	Removed
April 2013	44.6%	51.1%	47.9%	43.3%	Removed
April 2014	**Removed**	**Part 1 2 new items 51.6%**	**Part 2 2 edited items 53.4%**	**Part 3 3 edited items 43.1%**	**Part 4 6 (all) new items 28.6%**

Finally, similar to our observation made about Table 10, Table 14 shows that the Vocabulary, Grammar and Reading section had facility values that were very close to the values for that section of the test in the past, another indication that the unchanged parts of the test performed consistently this year with previous years.

Table 14: Average Scores by Part for Section 2 (reported as percent correct)

	Section 2: Vocabulary, Grammar and Reading		
TEST	**Part 6**	**Part 7**	**Part 8**
April 2010	56.4%	52.1%	40.6%
April 2011	56.8%	51.5%	41.9%
April 2012	60%	54.8%	51.9%
April 2013	58.9%	55.4%	50.9%
April 2014	**Part 5 59.5%**	**Part 6 54.9%**	**Part 7 49%**

E. Complete versus Partial Scores for End-of-Year FEPT

One of the issues the Assessments Committee has addressed since 2012 is how reliably the test has yielded complete scores at the end of the year for the Academic Office, which uses those scores to place students in English classes after their freshman year (Hull and Brennan, 2014, p. 56). Version 2.3 was a 54-minute test, and that length made it very difficult for teachers to administer the entire test in one 45-minute class at the end of the year. Because of that, teachers often administered the test over two classes, and that resulted in a number of incomplete scores for students due to attendance problems at the end of the year. In cases of partial scores for students, the Academic Office has to refer to the entrance FEPT score for those students, which may be an inaccurate measure of their ability at the end of the year. An additional problem was that teachers would lose one instructional class if they administered the test over two class periods.

Unfortunately, Table 15 shows an increase in the percentage of incomplete scores from 2013 to 2014. The figure is below the level reached when the test was a 54-minute test but higher than would be desirable. Because there will always be students at the end of the year who arrive too late after a test has started for the score to be counted or who are completely absent from a class for which a test is scheduled, it may be unrealistic to expect greater reductions in partial scores than what has been achieved the last two years. Busy with other priorities this past year, the committee was not able to implement another solution to this problem it had considered last year which was to schedule a make-up exam for students who missed the test in their Freshman English class at the end of the year. Considering that the rate in incomplete scores rose this year after such a significant reduction last year, the committee may want to try implementing a make-up end-of-year exam for January of 2015.

Table 15: Complete Versus Partial Scores for End-of-Year FEPT, 2010- 2014

	Number of Examinees	Number of Complete Scores	Number of Partial Scores	Percentage of Partial Scores
2010-11	1047	979	70	6.6%
2011-12	871	827	44	5%
2012-13	916	902	14	1.5%
2013-14	987	948	39	4%

F. From Version 2.6 to Version 2.7

In order to address the issue of how the FEPT can be improved for next year, it is necessary to review its strengths and weaknesses in light of the preceding review of its performance in 2014:

1. The test continues to distribute scores widely enough to divide the students into their Freshman English classes.
2. Although the degree to which the test yields consistent results with this test population certainly remains acceptable, its reliability measure did decline slightly in 2014 after another drop in 2013.
3. The test's ability to clearly distinguish among student proficiency levels as measured by item discrimination values also remains at an acceptable level. However, like the measurement of reliability, the average item discrimination level dropped in 2014 following another decline in 2013. The fall in 2014 was largely due to the introduction of a new monolog part.
4. The test continues to be at an appropriate overall level of difficulty for Asia University's entering student population. However, the disproportionately difficult level of the monolog part appears to have been the major reason why that part had poor discrimination values. Finally, the committee has more progress to make in balancing the difficulty levels of the two major sections of the test and in modifying the test so that examinees proceed from easier to more difficult items as they move from one part to another within each section of the test.

One approach the committee could take for next year would be to try to revise the newly introduced monolog part so that it is both at a more appropriate level of difficulty

and more effective at separating out student proficiency levels. This would require another round of item creation and pilot testing with follow-up review to create Version 2.7. Considering the time and resources required for such an undertaking, and after successive committees have not succeeded in developing an effective monolog part or replacement for the word discrimination part, it might be both more expedient and productive to return to the approach the committee took in 2012 when it scrutinized Version 2.3 of the test to develop a more condensed form of that test for Version 2.4. Rather than developing new items and having to carry out a pilot test to create Version 2.4, items with weak discrimination index values in Version 2.3 were removed to create a shorter version of the test that could be administered in a shorter period of time (Hull, 2012, pp. 9-10).

Following that kind of approach, the committee could remove the monolog part of Version 2.6 and re-introduce the word discrimination part previous to the 2013 Version 2.5 of the test after eliminating items that had weak discrimination values. Table 15 shows the 8 best performing items from the former word discrimination part in terms of discrimination index values. The average discrimination value for that part would be around .30 if those items were selected. That would surpass the .20 threshold for keeping items in the test and represent a significant improvement over the .11 obtained by the monolog part in Version 2.6 or the .18 that resulted for the word discrimination part of Version 2.5 (Hull and Brennan, p. 53, 2014).

TABLE 15: Discrimination Indexes for Modified Word Discrimination Part

	Item 1	Item 3	Item 4	Item 5	Item 6	Item 7	Item 10	Item 12	**AVERAGE**
April 2010	0.36	0.23	0,41	0.51	0.26	0.14	0.26	0.21	0.28
April 2011	0.37	0.27	0.40	0.52	0.24	0.22	0.25	0.21	0.31
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 9	Item 11	
April 2012	0.44	0.25	0.42	0.55	0.12	0.15	0.22	0.20	0.29
AVERAGE	0.39	0.25	0.41	0.53	0.21	0.17	0.24	0.21	**0.30**

Table 16 shows that the average facility value for those items would be .57. This is a level of difficulty that would make it easier than the other parts of the Listening section, and that would help the committee come closer to achieving its goal of having a Listening section that proceeds from easier to more difficult items.

TABLE 16: Facility Values for Modified Word Discrimination Part

	Item 1	Item 3	Item 4	Item 5	Item 6	Item 7	Item 10	Item 12	**AVERAGE**
April 2010	0.66	0.55	0.45	0.6	0.44	0.76	0.34	0.84	0.58
April 2011	0.58	0.54	0.39	0.57	0.45	0.79	0.38	0.84	0.57
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 9	Item 11	
April 2012	0.65	0.47	0.46	0.51	0.44	0.79	0.37	0.86	0.57
AVERAGE	0.63	0.52	0.43	0.56	0.44	0.78	0.36	0.85	**0.57**

These two changes alone have the potential to result in significant improvement in the performance of the FEPT and would not require a great deal of time and resources to carry out.

III. Conclusion

Over the last two years, the Assessments Committee has devoted a considerable amount of time to reviewing and analyzing the performance of the FEPT, creating new items and new parts for it and then conducting pilot tests with follow-up analysis and review to produce the two most recent versions of the test. Although the committee has learned a lot about testing in the process, not much actual progress has been made in the test itself.

The analysis here of the performance of Version 2.6 of the FEPT indicates that although some improvement was introduced to the test by adding a few items to the picture identification part and newly edited items to other parts, the major change introduced to the test, the creation of a new monolog part of the test, did not result in improved performance of the test. Actually, there were slight declines in the reliability of the test and in the ability of the test to discriminate among different levels of student ability. Fortunately, the amount of decline was not critical. More importantly, overall the test continues to perform at a very acceptable level and enables the Assessments Committee to successfully divide entering freshman students into their Freshman English classes with very little need for reassignment of students in classes after initial placement.

One very practical approach to improving the test for next year has been presented in this paper, and the scale of the work required for that approach should fit within the committee's time allowances. In short, the monolog part of the test could be

eliminated and eight better performing items from the word discrimination part of Versions 2.3 and 2.4 could be re-introduced. This would require re-editing and re-printing the paper form of the test and the re-recording of a limited number of items for the audio to create a new FEPT Version 2.7, but a time consuming pilot test would not need to be done. Based on the analysis we have presented, this revision should result in improved performance of the test. However, continued monitoring of the performance of the FEPT will be necessary to confirm whether any changes that are made actually do result in improvement or not.

At the same time, the committee has another direction for the placement test it would like to continue to work on. Following up on a conclusion that the committee made last year that it would be a good idea to explore commercial alternatives to the FEPT (Hull and Brennan, 2014, p. 59), the committee has recently received official permission from the publisher of one of the textbook series currently used in our Freshman English program to pilot use of their placement test with our students. We had hoped to start piloting this test last year but took longer obtaining the publisher's permission than we had anticipated.

The publisher's test is specifically designed to assign students to the appropriate textbook level in their series. However, because of the time constraints we have in CELE for entrance testing in April and end-of-year testing in January, we will need to remove some parts of the publisher's test to shorten it and conduct a pilot test to assess how well the test works with our test population. The publisher has also given the committee permission to combine items from our current entrance test with items from their test if that results in a better placement test for CELE. This kind of flexibility may prove very important as we pilot the publisher's test and review how well it fits our student population. The committee's current plan is to begin piloting the test in the second half of the 2014-2015 academic year and then again shortly after entering students are tested in the spring of 2015 in order to compare its strengths and weaknesses with the current FEPT. If the results of the pilot testing are favorable, it may be possible to implement the test from the spring of 2016.

This is an exciting development because a well-respected and well-established publisher of textbooks and tests has considerably greater resources and expertise to create

a placement test of higher professional quality than the CELE Assessments Committee at Asia University. Another benefit of being able to use a placement test of this sort is that the content of the test is directly related to the actual curriculum of the Freshman English program. This is important for at least two reasons. First, the test scores have the potential to result in better placement decisions. Second, although the current FEPT has undergone years of review and been modified numerous times, it has never been based on the Freshman English curriculum and therefore has not been a valid instrument to measure student improvement from the beginning of the year to the end of the year. This publisher's placement test has the potential to be such an instrument.

References

Alderson, C. J., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., and Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Forster, D. E., & Kearney, M. (1997). Writing the Freshman English Test (FEPT). *ELERI Journal*, 5, 144-157.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill, Inc.

Haladyna, T. M., and Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37-50.

Haladyna, T. M., Downing, S. M., and Rodriuez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–334.

Hughes, A. 2009. *Testing for language teachers*. Cambridge: Cambridge University Press.

Hull, J. (2013). Review and analysis of Asia University's 2012 Freshman English Placement Test, transition from version 2.3 to version 2.4. *CELE Journal*, *20*, 1-11.

Hull, J. (2012). Modifying Asia University's Freshman English Placement Test. *CELE Journal*, *20*, 1-11.

Hull, J. (2012). Results of the 2010-11 FEPT and TOEIC tests. CELE *Journal*, *20*, 34-38.

Richards, J. C., and Bohlke, D. (2012). *Four corners: 1*. Cambridge: Cambridge University Press.

Richards, J. C., and Bohlke, D. (2012). *Four corners: 2*. Cambridge: Cambridge University Press.